Informatics in Primary Care 2012;20:225–31

Refereed paper

Managing and exploiting routinely collected NHS data for research

Vasa Curcin PhD MSc BSc
Research Fellow, Department of Computing, Imperial College London, UK

Michael Soljak PhD
Clinical Research Fellow

Azeem Majeed PhD
Professor of Primary Care

Department of Primary Care and Public Health, Imperial College London, UK

ABSTRACT

Introduction Health research using routinely collected National Health Service (NHS) data derived from electronic health records (EHRs) and health service information systems has been growing in both importance and quantity. Wide population coverage and detailed patient-level information allow this data to be applied to a variety of research questions. However, the sensitivity, complexity and scale of such data also hamper researchers from fully exploiting this potential.

Objective Here, we establish the current challenges preventing researchers from making optimal use of the data sets at their disposal, on both the legislative and practical levels, and give recommendations as to how these challenges can be overcome.

Method A number of projects has recently been launched in the UK to address poor research data-management practices. Rapid Organisation of Healthcare Research Data (ROHRD) at Imperial College, London produced a useful prototype that provides local researchers with a one-stop index of available data sets together with relevant metadata.

Findings Increased transparency of data sets' availability and their provenance leads to better utilisation and facilitates compliance with regulatory requirements.

Discussion Research data resulting from NHS data is often not utilised fully, or is handled in a haphazard manner that prevents full auditability of the research. Furthermore, lack of informatics and data management skills in research teams act as a barrier to implementing more advanced practices, such as provenance capture and detailed, regularly updated, data management strategies. Only by a concerted effort at the levels of research organisations, funding bodies and publishers, can we achieve full transparency and reproducibility of the research.

Keywords: data governance, electronic health records, open data, provenance

What is known about the subject

- The changing nature of data used in research studies from primary care data requires a shift towards secure and traceable medical research.
- Current governance models and research practices in the UK are preventing optimal exploitation of available data resources.

What this paper adds

- The legislative framework in the UK is at the root of governance issues faced by the research teams.
- Multiple initiatives in government are aiming to open up the health data for research in both academia and industry.
- Home-grown research data management software is showing the way towards a more manageable and secure sharing of data.
- Provenance of source data and research results is key to maintaining traceability of research.

Introduction

The two core aspects of health research data management are governance (who can access the data and to what purpose) and provenance (where does the data come from, and how was it processed). Both permeate each stage of the clinical research process, from data collection, through cleaning, processing and analysis, to publication and beyond. Changes to the regulatory and software frameworks in research are currently being introduced to recognise this and bring closer the vision of secure and traceable medical research.

The need for standardisation in these two areas is receiving increased attention. The Royal Society's June 2012 report *Science as an Open Enterprise*, made a number of recommendations about encouraging the publishing and sharing of research data.[1] In particular, it stressed the need for capturing the provenance of research data outputs, in terms of authorship, links to relevant data sources and the data-processing history of the presented results. Kush *et al*[2] give an overview of current efforts in standardising research data from electronic health records (EHRs), and many journals are encouraging authors to publish their research data along with their paper.[3]

Governance of research data

Each research data set has associated with it its own set of information governance regulations, which vary depending on the type of data, the presence of consent, the relevant data controller and the parameters of the data collection. For example, the Hospital Episode Statistics (HES) protocol[4] stipulates that HES data must be held on isolated workstations, whereas anonymised patient-level data from National Clinical Audits (NCAs) may be held on secured internet-connected servers. Some data sources differentiate between confidential (patient-identifiable) data and sensitive data, with the latter typically consisting of patients' ethnicity, geographical information (sometimes including general practice location), political and religious views, and criminal records. However, the exact definition of these two classes of data is variable, even for the data sources with the same controller.[5]

Another anomaly is that geographic National Health Service (NHS) data collected for clinical or administrative purposes can be used without consent for clinical audit and service evaluation, but not always for research. However, most uses of this data are for observational research, often indistinguishable from service evaluation, but from a governance perspective this use is not differentiated from interventional research such as clinical trials. For example, general practices and NHS trusts are identified and published in NCAs and other data sources such as the Quality and Outcomes Framework (QOF), but are not identifiable in some research data sets derived from EHRs, preventing analysis of healthcare factors associated with patient outcomes.

There is also no uniform guidance on when in the analysis process the data may be removed from the required secure environments to researchers' desktops. Is it when it has been stripped of sensitive data items, or only when it has been aggregated? These unresolved issues increase the likelihood of researchers inadvertently breaching data protection policies. Conversely, they may also cause researchers to bury their data inside isolated data silos within their institutions, keeping it away not only from the outsiders, but also from their peers who might be authorised to view it and use it for their own research.[6]

The restrictive nature of the models that data controllers currently employ is a direct consequence of an ambiguous legislative framework. The Nuffield Trust has produced a useful summary of the background and implications of the current data protection models for health professionals and patients.[7] Although as late as mid-1990s it was commonly recognised that patient data should be freely available for research, introduction of the Data Protection Act in 1998 changed this. Despite provisions for secure processing of identifiable data for medical research, the exact definitions of 'secure' and 'medical research' were omitted, and many data controllers chose to adopt a strict interpretation of the rules that became known as 'consent or anonymise', whereby either consent should be obtained from each participant, or data was fully anonymised at source, effectively preventing any linkage with other data sources.

Provisions for allowing linkage of patient-identifiable data were introduced through Section 251 of the NHS Act of 2006, when it became apparent that key data sources such as cancer registrations would be compromised by more stringent legislation. Nonetheless, the process of obtaining the necessary permissions for specific projects is still complex and time-consuming, requiring application to the Ethics and Confidentiality Committee of the National Information Governance Board for Health and Social Care, whose functions will soon be taken over by the Health Research Authority (HRA). Organisations such as the NHS Information Centre for Health and Social Care (NHSIC) have developed a business model by which they handle this application process, acting as trusted third parties, or safe havens, for linkage and deliver the linked data to researchers.

Opening the research landscape in UK

The Open Data[8] initiative aims to make publicly generated data free and available to everyone, in useful formats, subject to proper attribution. One part of that vision, directly relevant to health data management, is that any published research study should be accompanied by the full data from which it was derived, thus enabling the reader to verify the results for themselves.

The government wants the United Kingdom (UK) to become a world leader in the use of public data to generate growth and expand knowledge.[9] The Chancellor of the Exchequer's 2011 Autumn Statement set out a range of Open Data measures to boost growth in UK life sciences by transforming access to health and care data.[10] In response to this, NHSIC released detailed general practice-level prescribing data online for the first time in September 2011[11] and in September 2012, it started a secure data linkage service to deliver data extracts, using linked health record data from primary and secondary care and other sources at an unidentifiable, individual level.[12]

In summer 2012, the Prime Minister announced that relaxing regulations on the collection and use of patient data would help the UK to become the best place in the world to conduct cutting-edge research.[13] This will include amending the NHS Constitution to enable patient data collected for clinical purposes to be used for research unless patients opt out. To that goal, a new secure data service, the Clinical Practice Research Datalink (CPRD), has been established jointly by the NIHR and the Medicines and Healthcare products Regulatory Agency (MHRA) to provide anonymised linked data from general practice, HES and other sources. The General Practice Extraction Service (GPES) has also announced that it will start offering data from general practices, starting in April 2013.[14] The data linkage with social survey data is helped by the UK Data Archive, which makes freely available to bona fide researchers person-level survey data obtained by informed consent. Similar efforts are also present in the USA, with federal government and private programmes that are dedicated to creating patient databases for research, such as the Million Veterans Program and the Kaiser Research Programme on Genes, Environment and Health.

Discovery and usage of research data

This wide availability of data brought to the forefront initiatives to establish data registries within institutions and encourage data sharing, by facilitating access and discovery. The Joint Information Systems Committee (JISC) has been particularly supportive of data-management initiatives in the UK, through its support of the Digital Curation Centre, which aims to build capacity, capability and skills for research data management, as well as establishing a range of projects in academic institutions under its Managing Research Data stream.

One example of a JISC-funded project in this area is the Rapid Organisation of Health Research Data (ROHRD) project,[15] which identified three key objectives for implementation of research data solutions:

- navigation of potentially vast data repositories within and outside their institutions to find the data that satisfies researchers' requirements
- quick and unambiguous retrieval of the data access policies attached to that data
- uniform and efficient process to obtain the data sets themselves.

ROHRD delivered a metadata model to support the first two objectives. The model, implemented as an ontology web language (OWL) ontology, uses the SNOMED-CT clinical vocabulary,[16] which is now a standard clinical coding system for the NHS, and captures contextual metadata about the data sets, including content description, governance restrictions and access procedure. An example of a metadata entry for one typical data set is given in Figure 1.

As a further step to support rapid visual querying of the data set properties, ROHRD has implemented a prototype web portal, linked from the metadata information that gives an interactive overview of age and gender breakdowns, diagnostic code prevalence and drug prescription frequencies, as shown in Figure 2. Combined with the contextual and data governance information, this framework gives researchers a mechanism for data discovery.

Other similar initiatives include the Oxford DataFlow[17] and BRISSkit[18] projects, which are creating a two-stage data management infrastructure to enable researchers to work with, annotate, publish and permanently store research data. The DataCite programme,[19] for which the British Library leads in the UK, allows researchers to create unique digital object identifiers (DOIs) for their data sets, so that they can be published in the same way as journal articles. For

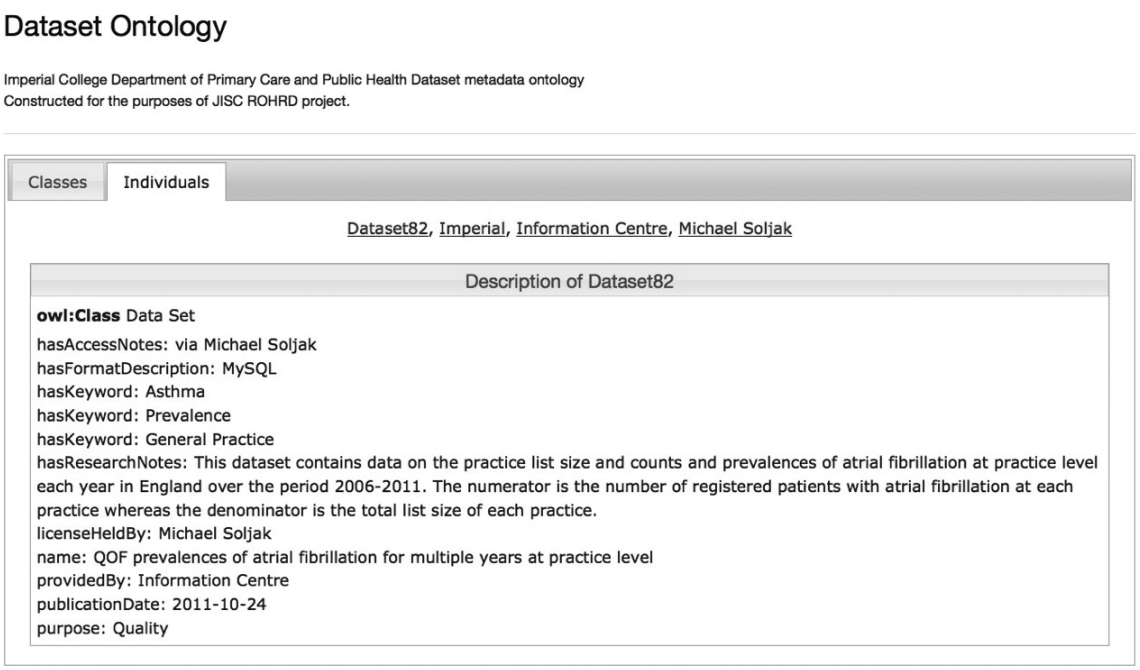

Figure 1 Entry for a research data set represented using Imperial's metadata ontology

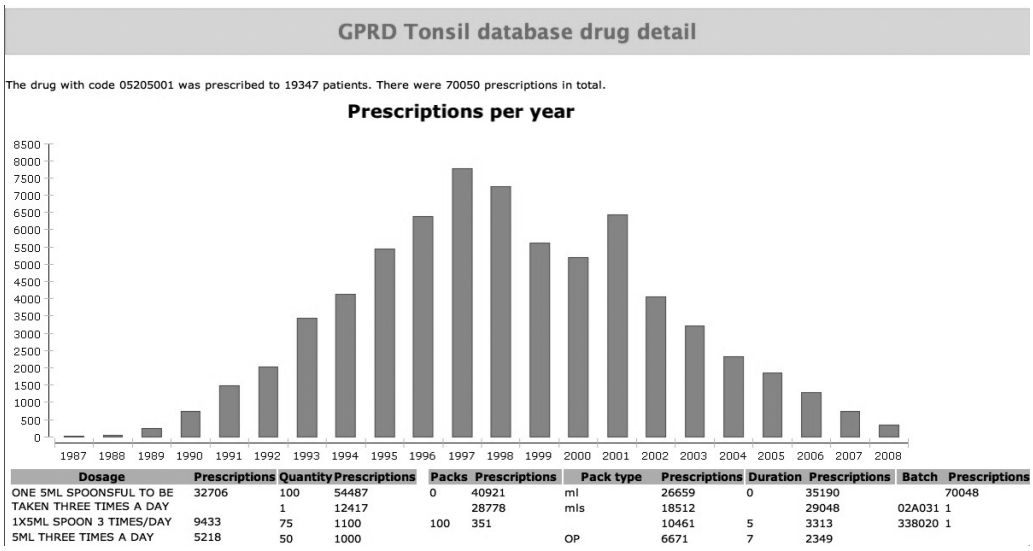

Figure 2 Visual representation of drug code metadata for a data source

example, following an outbreak of antibiotic-resistant *Escherichia coli* in Germany in 2011, BGI-Shenzhen included a DataCite DOI for the release of the genome into the public domain.[20]

Provenance of research data sets

Traceability and accountability of research data are essential components of data management in clinical research, with standards such as GxP (including Good Clinical Data Management Practice and Good Clinical Practice), CONSORT for trial reporting, and STROBE for reporting observational studies. Of particular interest is ADAM[21] produced by the Clinical Data Interchange Standards Consortium (CDISC), which documents each derived variable (treatment, outcome or covariate) used in clinical trial analysis data sets, to enable review and re-creation of published research.

The provenance of a piece of data refers to the knowledge about its origin, in terms of the entities and actors involved in its creation, e.g. data sources used,

operations carried out on them, together with the users enacting those operations. In research data management, provenance is concerned with both the study source data, and the resulting data sets produced. The former, *source provenance*, establishes where data originated, including the patient population profile and the governance restrictions, whereas the latter, *transformational provenance*, establishes the data linkage and transformations, e.g. joins and filters applied, performed on the source data to produce the study results. Whereas source provenance is typically manually entered and relatively small in size, transformational provenance, which aims to capture every operation applied to source data, can grow much larger, and has to be captured automatically, using software tools.

Making software systems provenance-aware enables investigation of data sources and services that produced a particular output from the software, together with the individuals who instigated the requests and received those outputs, to establish the exact lineage and assess that correct procedures were followed. The importance of provenance-aware infrastructures is reflected in the increasing number of biomedical research projects that include provenance components, such as EU FP7's TRANSFoRm[22] and EHR4CR.[23]

Provenance information is commonly represented as causal graphs establishing relationships between data entities, the processes that produced them and the agents (researchers, computer software) that performed those processes. The Open Provenance Model[24] (OPM) is a popular community standard for provenance description that is serving as a basis for the official World Wide Web Consortium PROV standard currently in development.[25] Provenance standards are closely related to semantic web technologies, and are based on uniform resource identifiers (URIs) for element identification, resource description framework (RDF) and OWL ontologies. Usage of semantic web technologies also enables the suitably annotated provenance graphs to be easily queried using standard medical ontologies, such as SNOMED-CT, ICD-10 or Read codes. However, there are no formal standards on the required level of provenance support in health research.

Discussion

The importance of research data management is increasingly recognised by both research funders and journal publishers. The EU Commission, the UK's Medical Research Council and the US National Institute of Health are just some of the funding bodies that now require research data management plans (RDMPs) as part of funding applications. A useful overview of the UK funders' requirements is maintained by the Digital Curation Centre.[26] Major publishers are also developing mechanisms, including data repositories and research data set citation indices, to provide full provenance of the data used for published research, ensuring traceability and reproducibility.

This change is mirrored within research institutions, which are slowly abandoning local data silos, and instead developing metadata registries, typically as internal web portals, to collect information about available data sets and their governance information, and to establish access procedures. Such efforts assist in both regulatory compliance and full exploitation of available data, but how these various initiatives will evolve, and if they will converge, remains unclear.

This provenance-gathering continues in the studies conducted with research data, with increasing numbers of software tools supporting automated provenance capture, thereby documenting the analysis steps, and facilitating complete research reporting. However, many research groups are still relying on non-standardised, fragmented software that is separated from the overarching data management strategy, disconnecting the findings from the processes that produced them.

Software packages are now available that allow research institutions to create their own data registries and provenance-aware infrastructures with little effort,[27,28] yet lack of research data management training, and of local data managers and informatics specialists in medical research groups are hindering progress. Key to overcoming this in a sustainable manner is the creation of comprehensive health informatics training programmes in the UK, including research data management, at undergraduate and postgraduate levels, focusing on data management, information governance, software architectures and semantic web technologies. Our key recommendations for achieving this are given in Table 1.

Medical research is moving towards full traceability, where data can be followed from its original collection, via anonymisation and linkage, to the analysis performed and the publication of results. Essential to this vision is the opening of healthcare data, through improved information governance frameworks that make use of the current advances in information technology. However, in order to sustain the changes indicated by the various projects in the area, there is a need for clearer and more consistent policies, more trained data managers, software architects and semantic web specialists in medical research groups.

Table 1 Addressing the challenges of health research data management

Challenge	Recommendation	Stakeholders
Insufficient incentive for researchers to publish datasets	Academic funders and institutions to add dataset citation indices to research excellence assessment, with clear mechanisms for referencing (e.g. DOIs)	Academia, government, publishers
Governance models outdated and too restrictive, with little or no audit of adherence	More devolved approval process for dataset usage needed, with proactive approach by the Health Research Authority, which is taking over from National Information Governance Board	Government, NHS
Lack of awareness of data available to researchers within institutions	Introduce metadata registries where users can find details on available data sets and their governance and provenance information	Academia, industry
Little or no provenance captured during data analysis	Increase usage of provenance-aware software tools and middleware in standard research practice, and incorporate it into publication requirements	Academia, industry, publishers
Poor data management and lack of coherent analytical software strategy	Better health informatics training and permanent data manager and software architect positions in health research groups	Academia, industry

CONTRIBUTORS AND SOURCES

Vasa Curcin is a Research Fellow at Department of Computing, Imperial College London, researching data provenance and analytical software architectures. He is the Scientific Manager of EU FP7 TRANSFoRm project and was an investigator on Imperial's ROHRD project. Michael Soljak is a public health researcher and consultant with a particular interest in information and intelligence, whose work has been acknowledged through involvement in national strategy in recent years. Azeem Majeed is Professor of Primary Care at Imperial College London, where he represents the Faculty of Medicine on the college's Research Data Management Board.

STATEMENT

FUNDING

The work in this paper has been partially funded by the JISC ROHRD project.

CONFLICTS OF INTEREST

All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years, no other

relationships or activities that could appear to have influenced the submitted work.

REFERENCES

1 The Royal Society. *Science as an Open Enterprise. The Royal Society Science Policy Centre report 02/12*. London: The Royal Society, 2012. royalsociety.org/policy/projects/science-public-enterprise/report/ (accessed 16 October 2012).
2 Kush RD. EHRs for clinical research. *AMIA Standards* Winter 2011–2012;2(2). www.amia.org/news-and-publications/volume-2-number-2/interoperability-review-2 (accessed 16 October 2012).
3 Groves T and Godlee F. Open science and reproducible research. *BMJ* 2012;344:e4383. www.bmj.com/content/344/bmj.e4383
4 Hospital Episode Statistics. *HES Protocol*. www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937 (accessed 16 October 2012).
5 Health and Social Care Information Centre. *Data Linkage Applications*. www.hscic.gov.uk/article/2184/Applications-and-Approvals (accessed 16 October 2012).
6 Weng C, Appelbaum P, Hripcsak G *et al*. Using EHRs to integrate research with patient care: promises and challenges. *Journal of the American Medical Informatics Association* 2012;19:684–7. Epub 2012 Apr 29.
7 Clark S and Weale A. *Access to Person-Level Data in Health Care: understanding information governance*. London: The Nuffield Trust, 2011. www.nuffieldtrust.org.uk/sites/files/nuffield/information_governance_in_health_-_research_report-_aug11.pdf
8 Panton Principles. *Open Data in Science*. pantonprinciples.org/ (accessed 16 October 2012).
9 Cabinet Office. *Transparency and Open Data Team UCO. Opening up government*, 2012. data.gov.uk/about-us (accessed 16 October 2012).
10 Cabinet Office. *Open Data Measures in the Autumn Statement 2011*. www.cabinetoffice.gov.uk/resource-library/open-data-measures-autumn-statement-2011 (accessed 16 October 2012).
11 NHS Information Centre for Health & Social Care. *Prescribing by GP Practice*, 09/2011 ed. http://www.hscic.gov.uk/gpprescribingdata (accessed 16 October 2012).
12 Department of Health. *The Government Plan for a Secure Data Service: strengthening the international competitiveness of UK life sciences research*. London: Department of Health, 2011. www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/documents/digitalasset/dh_131242.pdf (accessed 16 October 2012).
13 Collecting patient data will help UK become world leader in research, says Cameron. *BMJ* 2012;345 doi: 10.1136/bmj.e5285.
14 NHS Information Centre for Health & Social Care. *General Practice Extraction Service*. www.ic.nhs.uk/gpes (accessed 16 October 2012).
15 Imperial College London. *Rapid Organisation of Healthcare Research Data*. rapidhealthdata.wordpress.com/ (accessed 16 October 2012).
16 International Health Terminology Standards Development Organisation. *SNOMED Clinical Terms*. Copenghagen: International Health Terminology Standards Development Organisation, 2012. www.ihtsdo.org/snomed-ct/ (accessed 16 October 2012).
17 University of Oxford. *Data Flow*. www.dataflow.ox.ac.uk/ (accessed 16 October 2012).
18 University of Leicester. *BRISSkit*. http://www.brisskit.le.ac.uk/ (accessed 16 October 2012).
19 German National Library of Science and Technology. *DataCite*. http://www.datacite.org/ (accessed 16 October 2012).
20 Rohde H, Qin J, Cui Y *et al*. Open-source genomic analysis of shiga-toxin-producing *E. coli* O104:H4. *New England Journal of Medicine* 2011;365:718–24. dx.doi.org/10.1056/NEJMoa1107643 (accessed 16 October 2012).
21 Clinical Data Interchange Standards Consortium. *CDISC Analysis Data Model*. www.cdisc.org/adam (accessed 16 October 2012).
22 TRANSFoRm consortium. *TRANSFoRm: translational research and patient safety in Europe*. www.transformproject.eu (accessed 16 October 2012).
23 EHR4CR Consortium. *EHR4CR: electronic health records for clinical research*. www.ehr4cr.eu (accessed 16 October 2012).
24 Moreau L, Freire J, Futrelle J, McGrath R, Myers J and Paulson P. *The Open Provenance Model (Specification 1)*. University of Southampton. Report ePrint ID: 264979. 2007.
25 World Wide Web Consortium. *W3C PROV Model Primer*. www.w3.org/TR/prov-primer/ (accessed 16 October 2012).
26 Digital Curation Centre. *Overview of Funders' Data Policies*. www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies (accessed 16 October 2012).
27 Simmhan YL, Plale P and Gannon G. *A framework for collecting provenance in data-centric scientific workflows*. Proceedings of the 6th International Conference on Web Services, 2006, Chicago, pp. 427–36. doi: 10.1109/ICWS.2006.5.
28 Schlauch T and Schreiber A. *DataFinder – a scientific data management solution*. Proceedings of PV 2007. www.pv2007.dlr.de/Papers/Schlauch_DataFinder.pdf (accessed 23 November 2012).

ADDRESS FOR CORRESPONDENCE

Vasa Curcin
Department of Computing
Imperial College London
London SW7 2AZ
UK
Email: vasa.curcin@imperial.ac.uk

Accepted April 2013